CiteSeer 말뭉치를 이용한 과학기술 문헌의 주제 분석

(Topic Analysis of Science and Technology Articles using CiteSeer Corpus)

정 한 민 † 강 인 수 ††

(Hanmin Jung) (In-Su Kang)

성 원 경 †

(WonKyung Sung)

요 약 과학기술 분야는 매우 빠른 발전 속도를 보이며 세부 분야 간 융·복합 현상이 빈번하게 일어나는 특징을 가지고 있다. 과학기술정보 말뭉치로부터 상기 특성을 분석해 내는 작업은 연구 주제 추이를 분석하고 주제 간 연관 관계를 파악하기 위해 필요하다. 본 연구는 과학기술 분야 - 특히 정보기술(Information Technology) 분야 - 에서 광범위하게 활용되고 있는 Citeseer 말뭉치로부터 추출된 주제를 이용하여 다양한 주제 분석을 수행하는 방안을 보이는 것을 목표로 한다. 특히, 연구개발 전주기 지원 시스템인 OntoFrame에서 주제가 어떠한 역할을 할 수 있는지 사례를 통해 실증하고자 한다.

키워드 : 과학기술 문헌, 주제 추출, 주제 분석, CiteSeer 말뭉치, OntoFrame

Abstract There have been enormous technological advances in science & technology domain and frequent convergences between its sub-domains. Topic analysis with science & technology corpus is a key process to grasp topic trends and relations between topics. The main objective of this research is to show various analytic approaches with topics extracted from CiteSeer corpus, which is widely used in information technology domain. This paper will also show a case study of OntoFrame, an R&D support system developed by KISTI, to reveal the role of topics on the system.

Key words : Science & Technology Articles, Topic Extraction, Topic Analysis, CiteSeer Corpus, OntoFrame

1. 서 론

인터넷의 발달과 함께 접근 가능한 말뭉치가 전 분야에 걸쳐 급격히 늘어나고 있다. 특히, 과학기술 분야는 매우 빠른 발전 속도를 보이는 동시에 세부 분야 간 융·복합 현상도 빈번히 일어나는 특징을 가지고 있다. 과학기술정보 말뭉치로부터 상기 특성을 분석해 내는 작업은 연구 주제 추이를 분석하고 주제 간 연관 관계를 파악하는 것과 밀접한 관련이 있다. 또한, 주제의 효용성을 높일 수 있는 방안 마련이 필요하지만 아직까지 주제가 실제 응용 시스템에서 어떻게 활용될 수 있는지를 연계하여 연구하는 일이 드물다.

과학기술정보 말뭉치 중 OAI(Open Archives Initiative) 서비스를 통해 쉽게 접근할 수 있는 Citeseer (http://citeseer.ist.psu.edu/) 말뭉치는 수십 년에 걸쳐 70여 만 건 이상의 논문을 보유하고 있으며, Google Scholar (http://scholar.google.com/) 등과 함께 많은 연구자들이 이용하고 있다. 주제 분석의 효용성을 실증하기 위해서 원문을 포함한 대용량 말뭉치가 필요한데, 그런 측면에서도 적합한 말뭉치라 할 수 있다. 또한, 주제는 단편적인 측면에서의 관찰보다 시간 추이에 따른 분석을 통해 다양한 현상을 파악할 수 있는 특징이 있다는 점도 고려해야 한다.

본 연구는 대용량 말뭉치로부터의 주제 추출 및 분석을 통해 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시 등의 응용에 주제가 어떻게 활용될 수 있는 지와 연구개발 전주기 지원 시스템인 OntoFrame에서 주제가 어떠한 역할을 할 수 있는지를 사례를 통해 살펴보고자 한다. 2장에서는 말뭉치를 이용한 주제 관련 연구를 살펴보고, 3장에서는 본 연구에서 제시하는 주제 추출 및 분석 방법을 소개한다. 4장에서는 OntoFrame에의 실제 적용을 통해 주제 추출 및 분석의 필요성을 보이고자 한다.

2. 관련 연구

말뭉치를 이용한 연구는 문서 요약, 문서 분류, 정보 추출 등 다양한 분야에서 이루어져 왔다. 특히, 말뭉치

· 이 논문은 제34회 추계학술대회에서 '과학기술 문헌 기반 주제 분석'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : KISTI 정보서비스연구팀 책임연구원
jhm@kisti.re.kr
wksung@kisti.re.kr
†† 정 회 원 : 경성대학교 컴퓨터정보학부 교수
dbaisk@ks.ac.kr

논문접수 : 2007년 12월 7일
심사완료 : 2008년 4월 24일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제14권 제5호(2008.7)

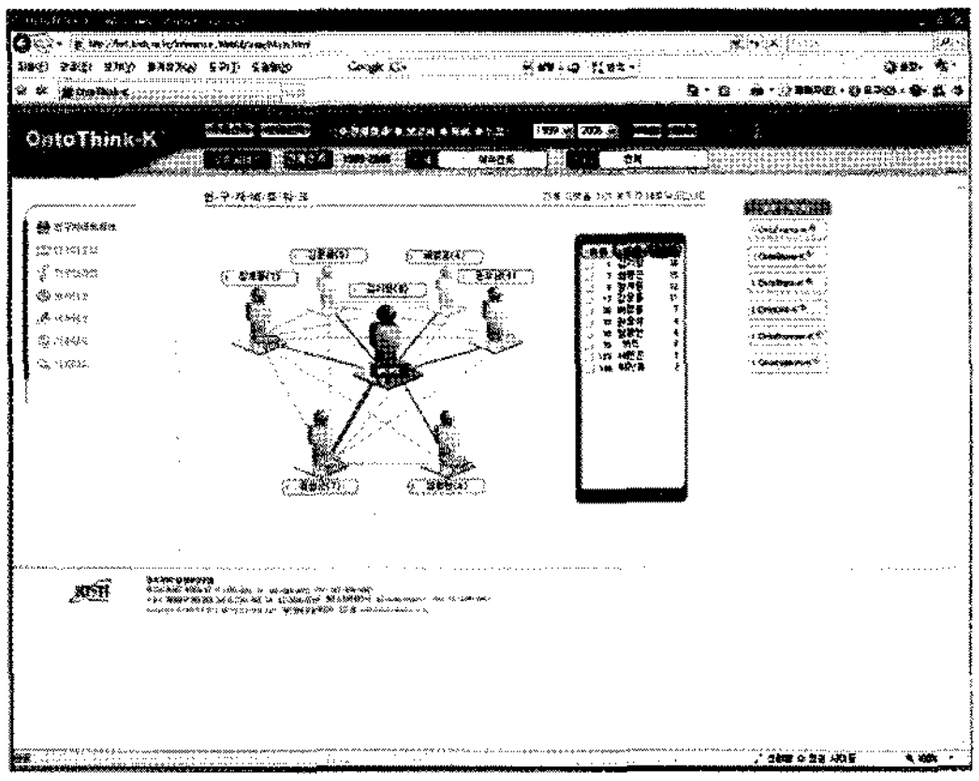

그림 1 2006년도 OntoFrame 화면 예
(주제 '에이전트'에 대한 연구자 네트워크 화면 예)

로부터 유용한 정보를 추출하여 활용하는 텍스트 마이닝(Text Mining)은 IT, BT 등 과학기술 여러 분야에서 적극 활용되고 있다. 주제와 관련된 연구 역시 활발히 이루어지고 있는데, 정보 추출 기법과 기계 학습 기법의 적용이 주를 이룬다.

주제와 관련된 연구는 주제를 단어나 구의 집합으로 정의하고 문서를 주제 집합으로 표현하는 방식과 주제를 단어나 구 자체로 정의하는 방식으로 나눌 수 있다. 전자에 속한 연구로는 [1]과 [2] 등이 있는데, [1]은 유사도에 기반을 두고 함께 뭉쳐진 중요 구들의 집합을 주제로 정의하고 웹 로그 분석을 통해 주제 추이를 시각화하여 보여준다. [2]는 저자, 주제, 단어를 연결하는 Author-Topic 모델과 Citeseer 말뭉치를 이용하여 주제 추이와 저자-주제-단어 연관 관계를 제시한다. 그렇지만, 단어 또는 구의 집합을 주제로 명시하기 때문에 주제명을 간접적으로 표현할 수밖에 없는 한계를 가진다. 또한, 각 주제를 표현하기 위한 단어 또는 구의 집합 크기가 주제에 따라 달라질 수 있으며, 특정 단어 또는 구가 여러 주제에 걸치게 되는 경우에 대해 고려해야 하는 등 복잡한 환경에 대한 제어가 쉽지 않다. 후자에 속한 연구는 단어 또는 구 자체를 주제로 정의하고 이들 간의 시간에 따른 추이를 비교하는 데 초점을 맞추고 있다. [3]과 [4]는 특정 분야에서 몇몇 주제들의 상대적 추이를 출현 빈도 등을 이용해 단순 비교함으로써 주제 추이를 직관적으로 보여준다. 그렇지만, 유사 주제간 관계를 파악하거나 문서를 주제 집합으로 표현하고 이 정보를 이용하는 방안을 제시하지 못하는 한계를 보인다.

[5]는 문서로부터 추출한 색인어를 시소러스 개념어와 매칭시킴으로써 해당 문서를 대표하는 주제와 분야 할당이 가능하도록 하는 자동 주제 및 분야할당 기법을 소개하고 있다. 이 방식은 주제를 통제할 수 있으며, 문서에 다중 주제를 할당함으로써 단편적인 문서 표현을 보완하여 융·복합적 성격의 문서 처리까지 가능하게 하는데 그 특징이 있다. 특히, 그림 1과 같이 연구개발 전주기 지원 시스템인 OntoFrame에 실제 적용되어 주제별 연구자 네트워크, 전문가 추천, 연구 성과 맵 등에 활용된다 [6]. 결국, 말뭉치로부터 획득한 주제를 분석하고 활용하기 위해서는 주제를 명확히 정의하고 응용 분야에 맞도록 모델링하여야 할 것이다.

3. 주제 추출 및 분석

본 장에서는 대용량 말뭉치로부터 추출된 전문용어 사전을 이용하여 문서의 주제를 결정하고 주제간 관계를 분석하는 방법을 설명한다. 그림 2는 워크플로우를 통해 이러한 과정을 설명하고 있다.

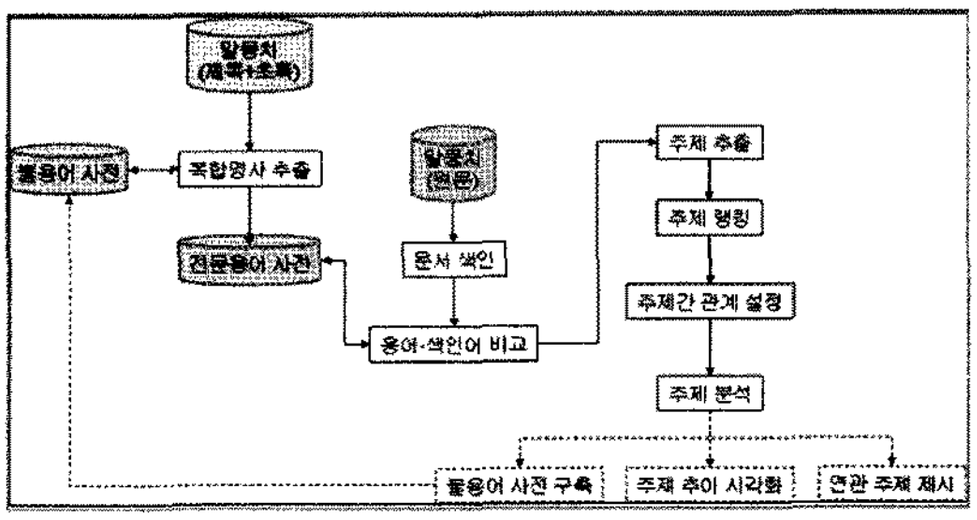

그림 2 주제 추출 및 분석 워크플로우

주제 추출 및 분석을 위해서는 말뭉치와 전문용어 사전이 필요하다. 말뭉치는 전문용어 사전 구축과 주제 추출을 위한 소스인데, 본 연구에서는 전문용어 사전 구축을 위해서는 제목과 초록으로 구성된 말뭉치를, 주제 추출을 위해서는 문서 단위로 구성된 원문 말뭉치를 사용한다. 전문용어 사전은 불용어를 최소한으로 포함하여야 하므로, 문서 내용을 잘 요약하고 있는 제목과 초록만을 이용해서 획득한다. 문서 색인은 정보 검색에서 검색 대상어를 결정하여 저장하는 과정으로, 본 연구에서는 통제 어휘를 선별하기 위한 전 단계(Pre-stage)로 이용한다. 우리는 용어-색인어 비교를 통해 선정된 통제 어휘를 주제라고 정의한다. 단일어는 용어 애매성과 비분야 처리 문제가 선결되어야 하기 때문에 편의상 복합명사만을 주제로 사용한다. 3.1절에서 기술한 방식으로 문서를 대표하는 주제들을 추출하여 순위화하는 과정까지 거치면 각 문서에 상위 N개(본 연구에서는 3개)의 주제가 할당된다. 2006년도 OntoFrame에서는 이 정보만을 활용하여 서비스를 구성하였으나, 2007년도 OntoFrame에서는 더 나아가 주제간 관계 설정 및 주제 분석을 통해 연관 주제를 제시하고 주제 추이를 시각화한다(4장 참조).

3.1 주제 추출

[5]는 한글 문서에 대해 시소러스 개념어-색인어 비교를 통한 주제 추출 및 순위화 알고리즘을 제시하였다. 본 연구에서는 대상 말뭉치가 영어로 구성된 대용량 Citeseer 말뭉치이므로 과학기술 분야의 영어 시소러스를 처리 범위(Coverage)를 보장하는 수준으로 구축하기 어려운 점을 고려하여 말뭉치로부터 추출된 전문용어 사전을 이용한다. 주제 추출을 위한 자질로는 문서로부터 추출된 색인어 목록, 말뭉치(제목+초록)로부터 추출된 전문용어 목록, 색인어 별 어휘 빈도(Term Frequency), 색인어와 매칭된 전문용어 별 어휘 빈도가 있다. 개념어 별 시소러스 깊이와 개념어 별 개념 패싯(Conceptual Facet)은 전문용어 사전이 시소러스와 같은 계층 구조를 가지고 있지 않아 이용하지 않는다.

Citeseer 말뭉치(제목+초록)로부터 추출된 전문용어 사전의 엔트리 수는 160,567개이며, 말뭉치 내에 포함된 문서 수는 127,029개(2000년 이후 논문)이다. 표 1은 주제 추출과 순위화를 거쳐 문서에 할당된 주제와 확률값을 보여준다.

표 1 문서에 할당된 상위 3개 주제와 확률값 예

문서 ID	주제1 ID	주제1 확률값	주제2 ID	주제2 확률값	주제3 ID	주제3 확률값
557	147284	0.076023	154374	0.052632	45438	0.052632
625	14818	0.092025	76675	0.08589	116922	0.079755
1043	120478	0.098131	102013	0.093458	102019	0.088785
1098	15382	0.105691	67173	0.044715	159989	0.044715
2903	47571	0.328358	66651	0.061834	11091	0.059701

주제간 관계 설정은 한 문서 내에서 선정된 상위 N개 주제를 상호 연결하는 방식으로 이루어진다. 공기(Coocurrence) 정보의 하위 집합이나, 문서를 대표할 수 있는 주제에 대해서만 주제간 관계를 설정하기 때문에 주제 대표성과 유의미한 주제 추이를 관찰하는데 도움이 된다. 예를 들어, 557번 문서에 상위 3개 주제를 할당한 경우, 147284-154374 ('Time Discretization'-'Value Problem'), 154374-45438 ('Value Problem'-'Error Bound'), 147284-45438 ('Time Discretization'-'Error Bound') 등 총 3개의 주제간 관계를 설정할 수 있다. 2000년 4,282개 문서에 대해 주제간 관계를 설정하면 전체 주제 6,765개 중 5,277개(78%)가 하나의 주제 네트워크로 연결된다(그림 3의 위쪽).

3.2 주제 분석

이는 주요 주제 허브(Topic Hub)에 해당하는 'Bayesian Net (ID: 11582)', 'Boundary Condition (ID: 14059)', 'Computer Science (ID: 24654)' 등이 대다수 주제를

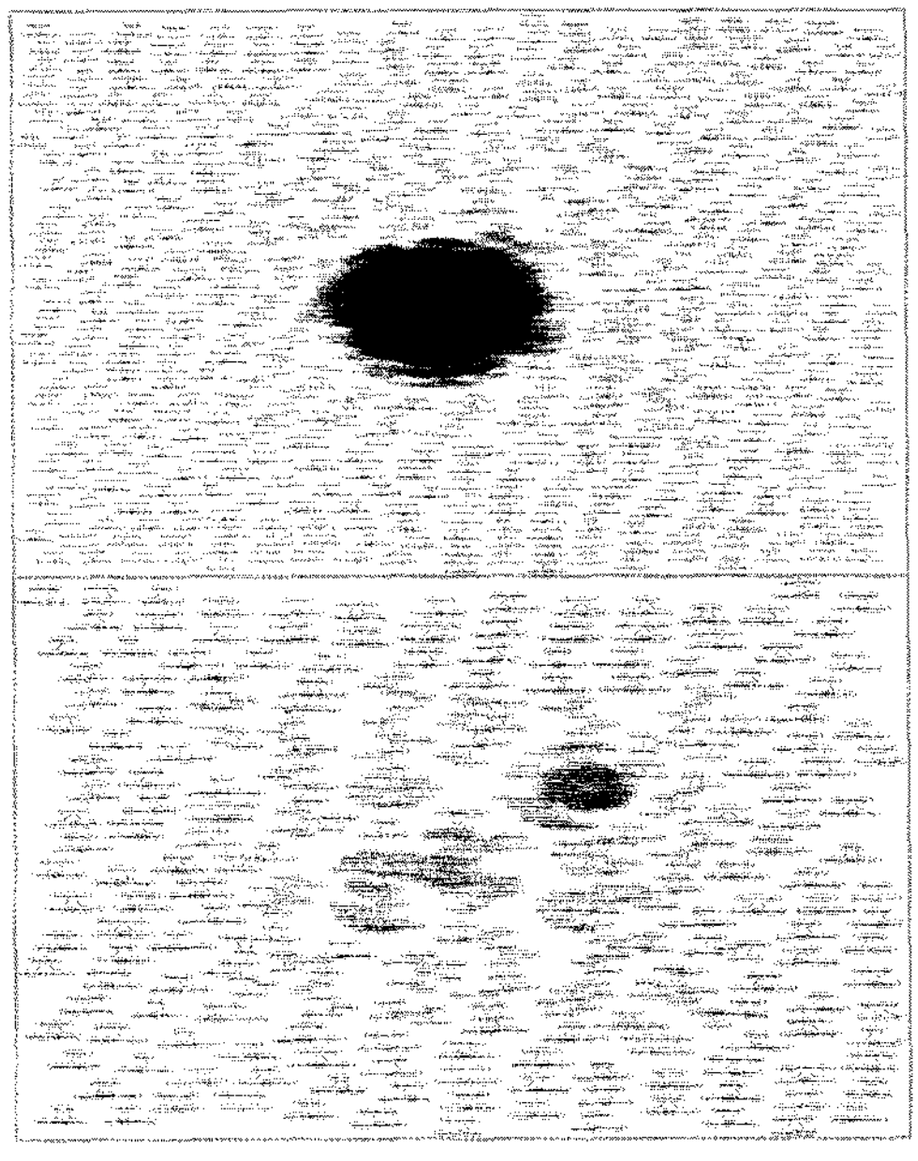

그림 3 주제간 관계 화면 예 (2000년, 2001년)

모으는 역할을 하기 때문이다. 그림 3은 2000년 문서 4,282개와 2001년 문서 318개에 대해 주제간 관계를 설정한 예를 보여주는데, 문서가 늘어날수록 주제 허브의 영향으로 주제간 연계가 2000년 말뭉치처럼 Big Head 구조를 가질 것으로 예측된다.

본 연구에서는 주제 분석을 3가지 관점에서 보고자 한다. 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시가 여기에 해당된다.

불용어 사전 구축을 위한 불용어 선정은 주제가 가지는 관계 수를 이용한다. 예를 들어, 2000~2002년 말뭉치에서 'The Set (ID: 146048)'의 경우 566개 주제와 관계를 가지며, 'This Paper (ID: 146400)'의 경우 143개 주제와 관계를 가지며, 'Data Set (ID: 32101)'의 경우 141개 주제와 관계를 가지는 등 주요 주제 허브에서 불용어 후보들이 나타난다. [5]에서도 DF (Document Frequency)가 극단적인 값(최소값과 최대값)을 가질수록 불용어일 확률이 높아진다는 유사한 실험 결과가 있다. 그렇지만, DF와 주제가 가지는 관계 수가 일치하는 것은 아니며, 상호보완적으로 적용할 필요가 있다(표 2 참조).

주제 추이 시각화는 특정 주제와 관계를 가지는 주제들이 시간 추이에 따라 변화하는 모습을 시각화시키고 관찰하는 것으로, 그림 4와 같은 Tag Cloud 등을 이용할 수 있다. 먼저 연도별로 말뭉치를 분리하고 각각의 말뭉치에서 관찰하고자 하는 주제와 직접적 관계를 가지는 주제들과 그 빈도를 추출한다. 시간 추이에 따라 새롭게 나타나는 주제를 파악함으로써 관찰 대상 주제

표 2 상위 10개 주제 허브에 대한 DF 값과 순위 (이탤릭체로 표시된 주제는 불용어)

주제	주제 허브 크기
The Set	566
This Paper	143
Data Set	141
Lower Bound	112
Computer Science	99
Neural Network	95
Markov Chain	61
Ion Problem	58
Operating System	57
Time Series	57

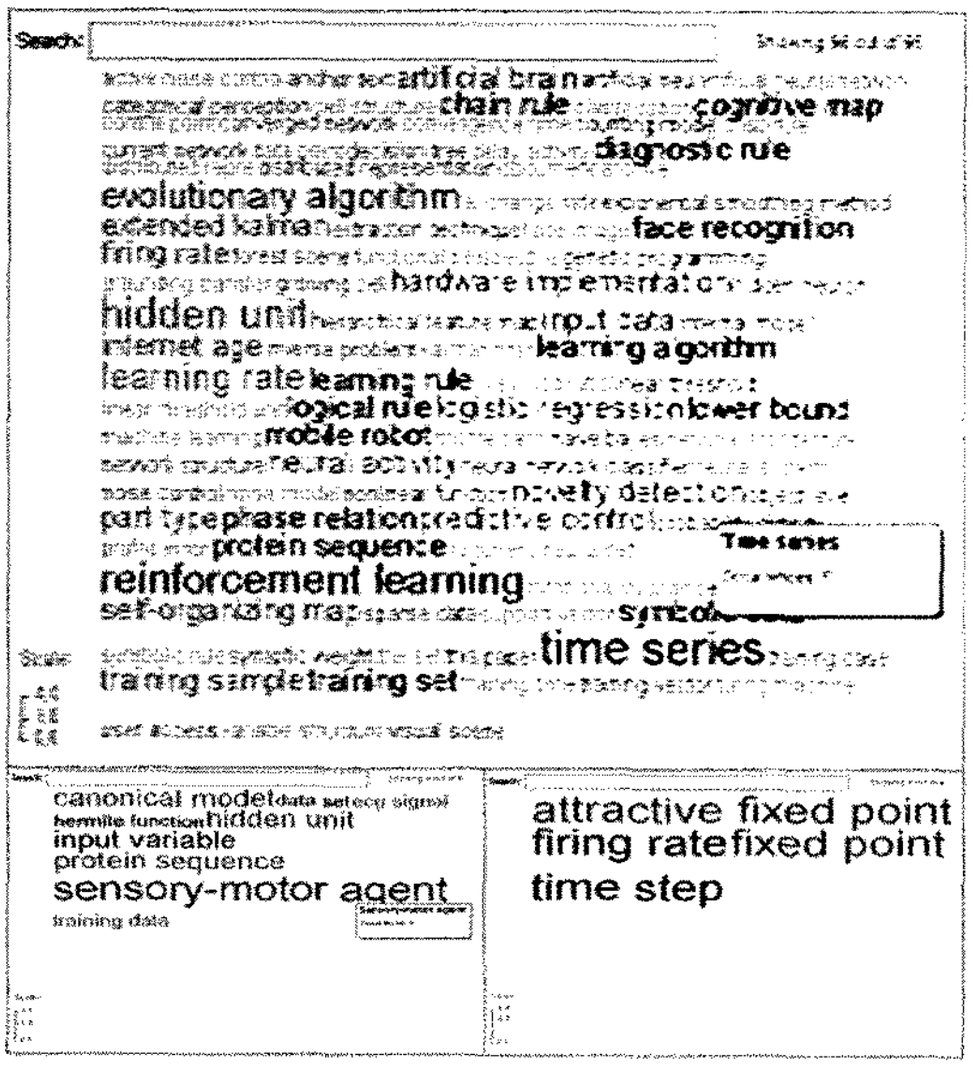

그림 4 연도별 연관 주제 Tag Cloud 화면 예 (2000~2002년 'Neural Net (work)' 연관 주제)[1)]

의 적용 영역 변화, 적용 방법 변화, 주제간 융·복합화 (예. 'Protein Structure'-'Metabolic Pathway'-'Information Extraction') 등을 확인할 수 있다. 예를 들어, 주제 'Neural Net (work)'의 경우, 2001년에 'Sensory- motor Agent', 'ECG Signal' 관련 응용 영역에 새롭게 적용되었으며, 'Hermite Function'에 기반한 신경망 모델이 제안되었다는 것과, 2002년에 'Fixed Point' 알고리즘이 신경망 모델 학습에 이용되었다는 사실 등을 알 수 있다.

연관 주제 제시는 특정 주제와 직접 또는 간접적으로 관련된 주제를 이용하여 주제간 관련도를 수치화시키는 방법이다. 연관 주제는 통합 검색 시 질의어로 입력된 주제와 연관된 주제를 보여주거나 질의어 확장에 이용할 수 있는데, 연관 주제에 가중치를 두면 가시화 범위나 확장 범위를 조절할 수 있다. 그림 5는 'Neural Net (work)'를 중심으로 한 연관 주제의 브라우징 화면을 보여준다. 'Hidden Unit', 'Protein Sequence', 'Training Data' 등 해당 주제와 직접 관계를 가지는 주제들은 직관적으로도 유사도가 크며, 'Spanning Tree', 'Wireless Network', 'Group Member' 등은 상대적으로 유사도가 떨어진다는 것을 알 수 있다. 의미적 거리는 해당 주제로부터 대상 주제까지의 떨어진 거리에 반비례하는 식으로 계산하여 활용할 수 있다.

4. 적용 사례

2006년도 OntoFrame은 국내 학술대회 논문을 서비스 대상으로 하였으며, 논문에서 추출된 주제들을 활용하여 논문을 분류하였다. 논문의 주제는 추론 엔진을 통해 해당 저자에게 전파(Propagation)되고, 저자에게 모인 주제는 재계산되어 저작 당시의 기관에 전파되는 방식으로 서비스 전반에 걸쳐 적용되었다. 특히, 주제를 시소러스 개념어로 통제함으로써 불용어와 분야 부적합어들을 배제할 수 있어 정교한 서비스 구현이 가능하였다. 주제별 세부 서비스를 살펴보면 다음과 같다.

1. 연구자 네트워크 (그림 1 참조): 주제별 연구자 그룹을 보여주고, 각 그룹에 대한 상세한 연구자 네트워크를 확인할 수 있다. 또한, 해당 주제의 대표적 전문가들을 성과 순으로 정렬하여 제시함으로써 전문가 추천을 지원한다.
2. 연구자 정보: 검색된 연구자의 대표 연구 주제들을 보여준다.
3. 연구 성과 맵: 광역 단위와 시군구 단위에서 주제별 성과 (논문, 특허, 보고서 등)를 보여준다. 해당 지역의 주제 정보는 논문, 연구자, 기관 순으로 전파된 주제를 재계산하여 얻어진다.
4. 성과 정보: 검색된 성과의 주제들을 상위 3개까지 보여준다. 3장에서 기술한 주제 추출 알고리즘을 통해 문서로부터 추출된 주제들로 해당 성과 (문서)를 표현한다.
5. 기관 정보: 검색된 기관의 대표 연구 주제들을 보여준다.

2007년도 OntoFrame은 논문, 인력, 주제, 출처 등 개체 별 서비스를 강화하는 방향으로 구현되었다. 특히, 주제 페이지를 별도로 구성하여, 해당 주제에 대한 연관 주제[2)] (그림 5 참조; 3.2절 참조), 주제 추이 (그림 6

1) Citeseer 말뭉치 내 2000~2002년 논문 중 4,971편의 원문으로부터 추출되고 가공된 주제 정보이며, http://www.tagcrowd.com/을 이용하여 시각화 함.

2) 2007년 OntoFrame에서는 시각화 문제로 인하여 목록 제시 화면으로 대체하였으며, 그림 5는 http://www.netvis.org/resources.php의 Social Network 시각화 도구를 이용하였음.

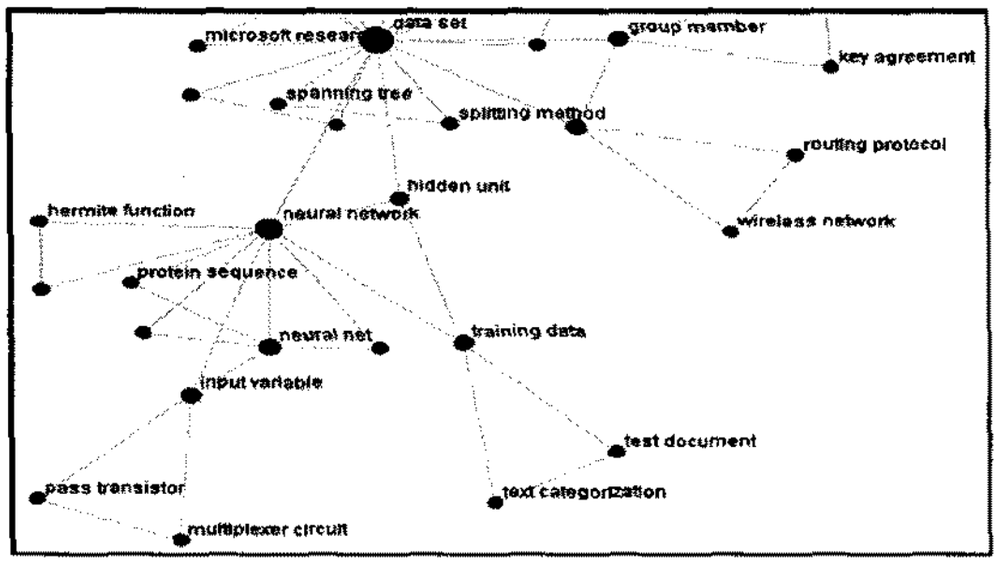

그림 5 연관 주제 브라우징 화면 예 (2001년 'Neural Net (work)' 연관 주제)

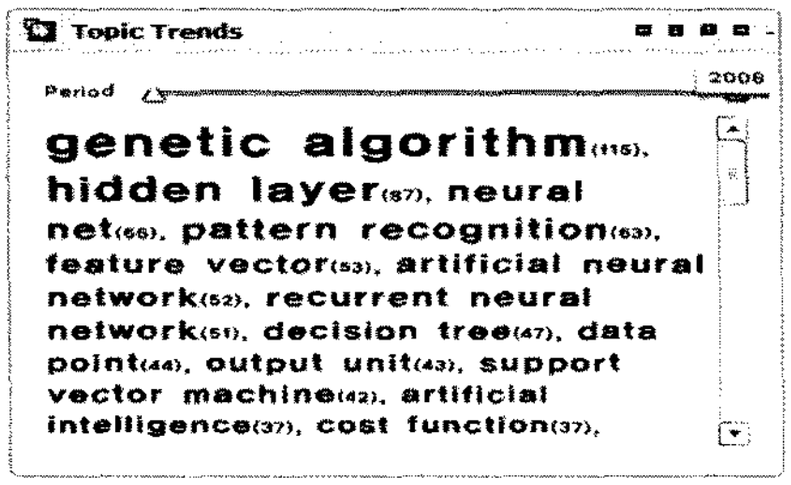

그림 6 주제 추이 OntoFrame 화면 예 (http://isrl.kisti.re.kr:8080/wsearch/search/main.jsp, 2000~2008년 'neural network')

참조), 논문, 인력, 기관, 위치, 연구자 네트워크, 학술대회, 도서 정보 등을 제시한다[7]. 연관 주제 및 주제 추이는 해당 주제 관련 방법론 진화, 응용 분야 확대 등의 환경적 변화를 인지하는데 도움을 준다.

5. 결 론

주제 추출 및 분석은 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시 등 다양한 응용에 활용될 수 있는 방법이다. 특히, 수작업으로 처리하기 어려운 대용량 말뭉치를 대상으로 하는 경우 자동적인 말뭉치 처리 기법을 사용할 수밖에 없다. 본 연구를 통해 대용량 Citeseer 말뭉치에서 주제를 추출하고 다양한 방식으로 분석하는 방법과 이를 연구개발 전주기 지원 시스템인 OntoFrame에서 활용한 사례를 소개하였다. 향후 연구는 다음 두 가지 작업을 포함할 것이다. 첫째는 불용어 처리를 통해 주제로서 적합한 용어를 선별하는 것이며, 둘째는 다양한 응용 분야에의 적용을 통해 본 연구에서 제시한 주제 분석 방안의 효용성을 보이는 것이다.

참 고 문 헌

[1] Glance, N., Hurst, M., and Tomokiyo, T., "BlogPulse: Automated Trend Discovery for Weblogs, Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics," 2004.

[2] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T., "Probabilistic Author-Topic Models for Information Discovery," Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[3] Lent, B., Agrawal, R., and Srikant, R., "Discovering Trends in Text Databases, Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining," 1997.

[4] Rajaraman K. and Tan, A., Topic Detection, "Tracking, and Trend Analysis Using Self-Organizing Neural Networks," Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2001.

[5] 정한민, 강인수, 성원경, "시소러스와 분야분류체계를 이용한 과학기술문헌에의 주제 및 분야할당", 제7회 한국언어정보학회 하계학술대회, 2006.

[6] Jung, H., Lee, M., Sung, W., and Park, D., "Semantic Web-Based Services for Supporting Voluntary Collaboration among Researchers Using an Information Dissemination Platform," Data Science Journal, Vol.6, 2007.

[7] Sung, W., Jung, H., Kim, P., Kang, I., Lee, S., Lee, M., Park, D., and Hahn S., "A Semantic Portal for Researchers Using OntoFrame," Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007.